

315 201 A  
08 322348



CERTIFICATION OF MAILING BY "EXPRESS MAIL"

Express Mail Label No. EF 942057005US

Date of Deposit: 13 October 1994

I hereby certify that this paper or fee is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 C.F.R. 1.10 on the date indicated above and is addressed to the Commissioner of Patents and Trademarks, Washington, D.C. 20231.

  
(Signature of person mailing paper or fee)

Stephen C. Malevitz

(Typed or printed name of person mailing paper or fee)

Case No. cbd1

## MOLECULAR TAGGING SYSTEM

### Field of the Invention

The invention relates generally to methods for identifying, sorting, and/or tracking molecules, especially polynucleotides, with oligonucleotide labels, and more particularly, to a method of sorting polynucleotides by specific hybridization to oligonucleotide tags.

### BACKGROUND

Specific hybridization of oligonucleotides and their analogs is a fundamental process that is employed in a wide variety of research, medical, and industrial applications, including the identification of disease-related polynucleotides in diagnostic assays, screening for clones of novel target polynucleotides, identification of specific polynucleotides in blots of mixtures of polynucleotides, amplification of specific target polynucleotides, therapeutic blocking of inappropriately expressed genes, DNA sequencing, and the like, e.g. Sambrook et al, Molecular Cloning: A Laboratory Manual, 2nd Edition (Cold Spring Harbor Laboratory, New York, 1989); Keller and Manak, DNA Probes, 2nd Edition (Stockton Press, New York, 1993); Milligan et al, J. Med. Chem., 36: 1923-1937 (1993); Drmanac et al, Science, 260: 1649-1652 (1993); Bains, J. DNA Sequencing and Mapping, 4: 143-150 (1993).

Specific hybridization has also been proposed as a method of tracking, retrieving, and identifying compounds labeled with oligonucleotide tags. For example, in multiplex DNA sequencing oligonucleotide tags are used to identify electrophoretically separated bands on a gel that consist of DNA fragments generated in the same sequencing reaction. In this way, DNA fragments from many sequencing reactions are separated on the same lane of a gel which is then blotted with separate solid phase materials on which the fragment bands from the separate sequencing reactions are visualized with oligonucleotide probes that specifically hybridize to complementary tags, Church et al, Science, 240: 185-188 (1988). Similar uses of oligonucleotide tags have also been proposed for identifying explosives, potential pollutants, such as crude oil, and currency for prevention and detection of counterfeiting, e.g. reviewed by Dollinger, pages 265-274 in Mullis et al, editors, The Polymerase Chain Reaction (Birkhauser, Boston, 1994). More recently, systems employing oligonucleotide tags have also been proposed as a means of manipulating and identifying individual molecules in complex combinatorial chemical libraries, for example, as an aid to screening such libraries for drug candidates, Brenner and Lerner, Proc. Natl. Acad. Sci., 89: 5381-5383 (1992); Alper, Science, 264: 1399-1401 (1994); and Needels et al, Proc. Natl. Acad. Sci., 90: 10700-10704 (1993).

The successful implementation of such tagging schemes depends in large part on the success in achieving specific hybridization between a tag and its complementary probe. That is, for an oligonucleotide tag to successfully identify a substance, the number of false positive and false negative signals must be minimized. Unfortunately, such spurious signals are not uncommon because base pairing and base stacking free energies vary widely among nucleotides in a duplex or triplex structure. For example, a duplex consisting of a repeated sequence of deoxyadenine (A) and thymidine (T) bound to its complement may have less stability than an equal-length duplex consisting of a repeated sequence of deoxyguanine (G) and deoxycytidine (C) bound to a partially complementary target containing a mismatch. Thus, if a desired compound from a large combinatorial chemical library were tagged with the former oligonucleotide, a significant possibility would exist that, under hybridization conditions designed to detect perfectly matched AT-rich

duplexes, undesired compounds labeled with the GC-rich oligonucleotide--even in a mismatched duplex--would be detected along with the perfectly matched duplexes consisting of the AT-rich tag. In the molecular tagging system proposed by Brenner et al (cited above), the related problem of mis-  
5 hybridizations of closely related tags was addressed by employing a so-called "commaless" code, which ensures that a probe out of register (or frame shifted) with respect to its complementary tag would result in a duplex with one or more mismatches for each of its five or more three-base words, or "codons."

10 Even though reagents, such as tetramethylammonium chloride, are available to negate base-specific stability differences of oligonucleotide duplexes, the effect of such reagents is often limited and their presence can be incompatible with, or render more difficult, further manipulations of the selected compounds, e.g. amplification by polymerase chain reaction  
15 (PCR), or the like.

Such problems have made the simultaneous use of multiple hybridization probes in the analysis of multiple or complex genetic loci, e.g. via multiplex PCR, reverse dot blotting, or the like, very difficult. As a result, direct sequencing of certain loci, e.g. HLA genes, has been promoted  
20 as a reliable alternative to indirect methods employing specific hybridization for the identification of genotypes, e.g. Gyllensten et al, Proc. Natl. Acad. Sci., 85: 7652-7656 (1988).

The ability to sort cloned and identically tagged DNA fragments onto distinct solid phase supports would facilitate such sequencing,  
25 particularly when coupled with a non gel-based sequencing methodology simultaneously applicable to many samples in parallel.

In view of the above, it would be useful if there were available an oligonucleotide-based tagging system which provided a large repertoire of tags, but which also minimized the occurrence of false positive and false  
30 negative signals without the need to employ special reagents for altering natural base pairing and base stacking free energy differences. Such a tagging system would find applications in many areas, including construction and use of combinatorial chemical libraries, large-scale mapping and sequencing of DNA, genetic identification, medical  
35 diagnostics, and the like.

### Summary of the Invention

The invention provides a method of tracking, identifying, and/or sorting classes or subpopulations of molecules by the use of oligonucleotide tags. An oligonucleotide tag of the invention consists of a plurality of subunits, each subunit consisting of an oligonucleotide of 3 to 6 nucleotides in length. Subunits of an oligonucleotide tag are selected from a minimally cross-hybridizing set. In such a set, a duplex or triplex consisting of a subunit of the set and the complement of any other subunit of the set contains at least two mismatches. In other words, a subunit of a minimally cross-hybridizing set at best forms a duplex or triplex having two mismatches with the complement of any other subunit of the same set. The number of oligonucleotide tags available in a particular embodiment depends on the number of subunits per tag and on the length of the subunit. The number is generally much less than the number of all possible sequences the length of the tag, which for a tag  $n$  nucleotides long would be  $4^n$ .

In one aspect of the invention, complements of oligonucleotide tags attached to a solid phase support are used to sort polynucleotides from a mixture of polynucleotides each containing a tag. In this embodiment, complements of the oligonucleotide tags are synthesized on the surface of a solid phase support, such as a microscopic bead or a specific location on an array of synthesis locations on a single support, such that populations of identical sequences are produced in specific regions. That is, the surface of each support, in the case of a bead, or of each region, in the case of an array, is derivatized by only one type of complement which has a particular sequence. The population of such beads or regions contains a repertoire of complements with distinct sequences, the size of the repertoire depending on the number of subunits per oligonucleotide tag and the length of the subunits employed. Similarly, the polynucleotides to be sorted each comprises an oligonucleotide tag in the repertoire, such that identical polynucleotides have the same tag and different polynucleotides have different tags. Thus, when the populations of supports and polynucleotides are mixed under conditions which permit specific hybridization of the oligonucleotide tags with their respective complements, subpopulations of identical polynucleotides are sorted onto particular beads or regions. The

subpopulations of polynucleotides can then be manipulated on the solid phase support by micro-biochemical techniques.

Generally, the method of the invention comprises the following steps: (a) attaching an oligonucleotide tag from a repertoire of tags to each  
5 molecule in a population of molecules (i) such that the same molecules or same subpopulation of molecules in the population have the same oligonucleotide tag attached and different molecules or different subpopulations of molecules in the population have different oligonucleotide tags attached and (ii) such that each oligonucleotide tag  
10 from the repertoire comprises a plurality of subunits and each subunit of the plurality consists of an oligonucleotide having a length from three to six nucleotides or from three to six basepairs, the subunits being selected from a minimally cross-hybridizing set; and (b) sorting the molecules or subpopulations of molecules of the population by specifically hybridizing  
15 the oligonucleotide tags with their respective complements.

Preferably, every subunit within a given minimally cross-hybridizing set has the same percentage composition of nucleotide types. For example, in one embodiment where subunits are 3-mers, each subunit may consist of  
20 66% A or T and 34% G or C. In this way, the stability of perfectly matched duplexes between every subunit and its complement is approximately equal.

An important aspect of the invention is the use of the oligonucleotide tags to sort polynucleotides for parallel sequence determination. This aspect of the invention comprises the following steps, (a) target polynucleotides are inserted into a conventional cloning vector, such as  
25 M13, or the like, that contains an oligonucleotide tag from a repertoire so that a library of target polynucleotide-tag conjugates is formed; (b) a sample of vectors is taken from the library and the individual vectors of the sample are amplified; (c) the target polynucleotide-tag conjugates are excised from the vectors and mixed with a repertoire of tag complements attached to  
30 solid phase supports under conditions that promote specific hybridization, thereby causing the polynucleotides to be sorted to the solid phase supports of the tag complements; and (d) separately determining the sequences of the polynucleotides attached to the solid phase supports. Preferably, the sequences of the target polynucleotides are determined by a process of  
35 stepwise ligation and cleavage described below, or like process not requiring electrophoretic separation of closely-sized DNA fragments. More

preferably, the solid phase supports are microparticles, such as controlled-pore glass (CPG), which are deposited on a flat surface, such as a glass microscope slide, for parallel, or simultaneous, sequencing.

The present invention overcomes a key deficiency of current methods of tagging or labeling molecules with oligonucleotides: By coding the sequences of the tags in accordance with the invention, the stability of any mismatched duplex or triplex between a tag and a complement to another tag is far lower than that of any perfectly matched duplex between the tag and its own complement. Thus, the problem of incorrect sorting because of mismatch duplexes of GC-rich tags being more stable than perfectly matched AT-rich tags is eliminated.

When used in combination with solid phase supports, such as microscopic beads, the present invention provides a readily automated system for manipulating and sorting polynucleotides, particularly useful in large-scale parallel operations, such as large-scale DNA sequencing, wherein many target polynucleotides or many segments of a single target polynucleotide are sequenced simultaneously.

#### Brief Description of the Drawings

Figure 1 is a flow chart illustrating a general algorithm for generating minimally cross-hybridizing sets.

#### Definitions

"Complement" or "tag complement" as used herein in reference to oligonucleotide tags refers to an oligonucleotide to which a oligonucleotide tag specifically hybridizes to form a perfectly matched duplex or triplex. In embodiments where specific hybridization results in a triplex, the oligonucleotide tag may be selected to be either double stranded or single stranded. Thus, where triplexes are formed, the term "complement" is meant to encompass either a double stranded complement of a single stranded oligonucleotide tag or a single stranded complement of a double stranded oligonucleotide tag.

The term "oligonucleotide" as used herein includes linear oligomers of natural or modified monomers or linkages, including deoxyribonucleosides, ribonucleosides,  $\alpha$ -anomeric forms thereof, polyamide nucleic acids, and the like, capable of specifically binding to a target polynucleotide by way of a regular

pattern of monomer-to-monomer interactions, such as Watson-Crick type of base pairing, base stacking, Hoogsteen or reverse Hoogsteen types of base pairing, or the like. Usually monomers are linked by phosphodiester bonds or analogs thereof to form oligonucleotides ranging in size from a few monomeric units, e.g. 3-4, to  
5 several tens of monomeric units. Whenever an oligonucleotide is represented by a sequence of letters, such as "ATGCCTG," it will be understood that the nucleotides are in 5'→3' order from left to right and that "A" denotes deoxyadenosine, "C" denotes deoxycytidine, "G" denotes deoxyguanosine, and "T" denotes thymidine, unless otherwise noted. Analogs of phosphodiester linkages  
10 include phosphorothioate, phosphorodithioate, phosphoranilidate, phosphoramidate, and the like.

"Perfectly matched" in reference to a duplex means that the poly- or oligonucleotide strands making up the duplex form a double stranded structure with one other such that every nucleotide in each strand undergoes  
15 Watson-Crick basepairing with a nucleotide in the other strand. The term also comprehends the pairing of nucleoside analogs, such as deoxyinosine, nucleosides with 2-aminopurine bases, and the like, that may be employed. In reference to a triplex, the term means that the triplex consists of a perfectly matched duplex and a third strand in which every nucleotide  
20 undergoes Hoogsteen or reverse Hoogsteen association with a basepair of the perfectly matched duplex. Conversely, a "mismatch" in a duplex between a tag and an oligonucleotide means that a pair or triplet of nucleotides in the duplex or triplex fails to undergo Watson-Crick and/or Hoogsteen and/or reverse Hoogsteen bonding.

25 As used herein, "nucleoside" includes the natural nucleosides, including 2'-deoxy and 2'-hydroxyl forms, e.g. as described in Kornberg and Baker, DNA Replication, 2nd Ed. (Freeman, San Francisco, 1992). "Analog" in reference to nucleosides includes synthetic nucleosides having modified base moieties and/or modified sugar moieties, e.g. described by  
30 Scheit, Nucleotide Analogs (John Wiley, New York, 1980); Uhlman and Peyman, Chemical Reviews, 90: 543-584 (1990), or the like, with the only proviso that they are capable of specific hybridization. Such analogs include synthetic nucleosides designed to enhance binding properties, reduce degeneracy, increase specificity, and the like.

### Detailed Description of the Invention

The invention provides a method of labeling and sorting molecules, particularly polynucleotides, by the use of oligonucleotide tags. The oligonucleotide tags of the invention comprise a plurality of "words" or subunits selected from minimally cross-hybridizing sets of subunits. Subunits of such sets cannot form a duplex or triplex with the complement of another subunit of the same set with less than two mismatched nucleotides. Thus, the sequences of any two oligonucleotide tags of a repertoire that form duplexes will never be "closer" than differing by two nucleotides. In particular embodiments, sequences of any two oligonucleotide tags of a repertoire can be even "further" apart, e.g. by designing a minimally cross-hybridizing set such that subunits cannot form a duplex with the complement of another subunit of the same set with less than three mismatched nucleotides. The invention is particularly useful in labeling and sorting polynucleotides for parallel operations, such as sequencing or fingerprinting.

### Constructing Oligonucleotide Tags from Minimally Cross-Hybridizing Sets of Subunits

The nucleotide sequences of the subunits for any minimally cross-hybridizing set are conveniently enumerated by simple computer programs following the general algorithm illustrated in Fig. 1, and as exemplified by program minhx whose source code is listed in Appendix I. Minhx computes all minimally cross-hybridizing sets having subunits composed of three kinds of nucleotides and having length of four.

The algorithm of Fig. 1 is implemented by first defining the characteristic of the subunits of the minimally cross-hybridizing set, i.e. length, number of base differences between members, and composition, e.g. do they consist of two, three, or four kinds of bases. A table  $M_n$ ,  $n=1$ , is generated (100) that consists of all possible sequences of a given length and composition. An initial subunit  $S_1$  is selected and compared (120) with successive subunits  $S_i$  for  $i=n+1$  to the end of the table. Whenever a successive subunit has the required number of mismatches to be a member of the minimally cross-hybridizing set, it is saved in a new table  $M_{n+1}$  (125), that also contains subunits previously selected in prior passes through step 120. For example, in the first set of comparisons,  $M_2$  will



contain  $S_1$ ; in the second set of comparisons,  $M_3$  will contain  $S_1$  and  $S_2$ ; in the third set of comparisons,  $M_4$  will contain  $S_1$ ,  $S_2$ , and  $S_3$ ; and so on. Similarly, comparisons in table  $M_j$  will be between  $S_j$  and all successive subunits in  $M_j$ . Note that each successive table  $M_{n+1}$  is smaller than its predecessors as subunits are eliminated in successive passes through step 130. After every subunit of table  $M_n$  has been compared (140) the old table is replaced by the new table  $M_{n+1}$ , and the next round of comparisons are begun. The process stops (160) when a table  $M_n$  is reached that contains no successive subunits to compare to the selected subunit  $S_i$ , i.e.

10  $M_n = M_{n+1}$ .

A preferred embodiment of minimally cross-hybridizing sets are those whose subunits are made up of three of the four natural nucleotides. As will be discussed more fully below, the absence of one type of nucleotide in the oligonucleotide tags permits target polynucleotides to be loaded onto solid phase supports by use of the 5'→3' exonuclease activity of a DNA polymerase. The following is an exemplary minimally cross-hybridizing set of subunits each comprising four nucleotides selected from the group consisting of A, G, and T:

20

Table I

Word:	$w_1$	$w_2$	$w_3$	$w_4$
Sequence:	GATT	TGAT	TAGA	TTTG

Word:	$w_5$	$w_6$	$w_7$	$w_8$
Sequence:	GTAA	AGTA	ATGT	AAAG

25

In this set, each member would form a duplex having three mismatched bases with the complement of every other member.

Further exemplary minimally cross-hybridizing sets are listed below in Table I. Clearly, additional sets can be generated by substituting different

groups of nucleotides, or by using subsets of known minimally cross-hybridizing sets.

Table II

5 Exemplary Minimally Cross-Hybridizing Sets of 4-mer Subunits

CATT	ACCC	AAAC	AAAG	AACA	AACG
CTAA	AGGG	ACCA	ACCA	ACAC	ACAA
TCAT	CACG	AGGG	AGGC	AGGG	AGGC
ACTA	CCGA	CACG	CACC	CAAG	CAAC
TACA	CGAC	CCGC	CCGG	CCGC	CCGG
TTTC	GAGC	CGAA	CGAA	CGCA	CGCA
ATCT	GCAG	GAGA	GAGA	GAGA	GAGA
AAAC	GGCA	GCAG	GCAC	GCCG	GCCC
	AAAA	GGCC	GGCG	GGAC	GGAG
AAGA	AAGC	AAGG	ACAG	ACCG	ACGA
ACAC	ACAA	ACAA	AACA	AAAA	AAAC
AGCG	AGCG	AGCC	AGGC	AGGC	AGCG
CAAG	CAAG	CAAC	CAAC	CACC	CACA
CCCA	CCCC	CCCG	CCGA	CCGA	CCAG
CGGC	CGGA	CGGA	CGCG	CGAG	CGGC
GACC	GACA	GACA	GAGG	GAGG	GAGG
GCGG	GCGG	GCGC	GCCC	GCAC	GCCC
GGAA	GGAC	GGAG	GGAA	GGCA	GGAA

The oligonucleotide tags of the invention and their complements are conveniently synthesized on an automated DNA synthesizer, e.g. an  
10 Applied Biosystems, Inc. (Foster City, California) model 392 or 394 DNA/RNA Synthesizer, using standard chemistries, such as phosphoramidite chemistry, e.g. disclosed in the following references: Beaucage and Iyer, Tetrahedron, 48: 2223-2311 (1992); Molko et al, U.S. patent 4,980,460; Koster et al, U.S. patent 4,725,677; Caruthers et al, U.S.  
15 patents 4,415,732; 4,458,066; and 4,973,679; and the like. Alternative chemistries, e.g. resulting in non-natural backbone groups, such as phosphorothioate, phosphoramidate, and the like, may also be employed provided that the resulting oligonucleotides are capable of specific hybridization.

20 When microparticles are used as supports, repertoires of oligonucleotide tags and tag complements are preferably generated by

subunit-wise synthesis via "split and mix" techniques, e.g. as disclosed in Shortle et al, International patent application PCT/US93/03418. Briefly, the basic unit of the synthesis is a subunit of the oligonucleotide tag. Preferably, phosphoramidite chemistry is used and 3' phosphoramidite oligonucleotides are prepared for each subunit in a minimally cross-hybridizing set, e.g. for the set first listed above, there would be eight 4-mer 3'-phosphoramidites. Synthesis proceeds as disclosed by Shortle et al or in direct analogy with the techniques employed to generate diverse oligonucleotide libraries using nucleosidic monomers, e.g. as disclosed in Telenius et al, Genomics, 13: 718-725 (1992); Welsh et al, Nucleic Acids Research, 19: 5275-5279 (1991); Grothues et al, Nucleic Acids Research, 21: 1321-1322 (1993); Hartley, European patent application 90304496.4; Lam et al, Nature, 354: 82-84 (1991); Zuckerman et al, Int. J. Pept. Protein Research, 40: 498-507 (1992); and the like. Generally, these techniques simply call for the application of mixtures of the activated monomers to the growing oligonucleotide during the coupling steps.

Double stranded forms of tags are made by separately synthesizing the complementary strands followed by mixing under conditions that permit duplex formation. Such duplex tags may then be inserted into cloning vectors along with target polynucleotides for sorting and manipulation of the target polynucleotide in accordance with the invention.

In embodiments where specific hybridization occurs via triplex formation, coding of tag sequences follows the same principles as for duplex-forming tags; however, there are further constraints on the selection of subunit sequences. Generally, third strand association via Hoogsteen type of binding is most stable along homopyrimidine-homopurine tracks in a double stranded target. Usually, base triplets form in T-A\*T or C-G\*C motifs (where "-" indicates Watson-Crick pairing and "\*" indicates Hoogsteen type of binding); however, other motifs are also possible. For example, Hoogsteen base pairing permits parallel and antiparallel orientations between the third strand (the Hoogsteen strand) and the purine-rich strand of the duplex to which the third strand binds, depending on conditions and the composition of the strands. There is extensive guidance in the literature for selecting appropriate sequences, orientation, conditions, nucleoside type (e.g. whether ribose or deoxyribose nucleosides are employed), base modifications (e.g. methylated cytosine, and the like) in

order to maximize, or otherwise regulate, triplex stability as desired in particular embodiments, e.g. Roberts et al, Proc. Natl. Acad. Sci., 88: 9397-9401 (1991); Roberts et al, Science, 258: 1463-1466 (1992); Distefano et al, Proc. Natl. Acad. Sci., 90: 1179-1183 (1993); Mergny et al,  
5 Biochemistry, 30: 9791-9798 (1991); Cheng et al, J. Am. Chem. Soc., 114: 4465-4474 (1992); Beal and Dervan, Nucleic Acids Research, 20: 2773-2776 (1992); Beal and Dervan, J. Am. Chem. Soc., 114: 4976-4982 (1992); Giovannangeli et al, Proc. Natl. Acad. Sci., 89: 8631-8635 (1992); Moser and Dervan, Science, 238: 645-650 (1987); McShan et al, J. Biol. Chem.,  
10 267:5712-5721 (1992); Yoon et al, Proc. Natl. Acad. Sci., 89: 3840-3844 (1992); Blume et al, Nucleic Acids Research, 20: 1777-1784 (1992); Thuong and Helene, Angew. Chem. Int. Ed. Engl. 32: 666-690 (1993); and the like. Conditions for annealing single-stranded or duplex tags to their single-stranded or duplex complements are well known, e.g. Ji et al, Anal.  
15 Chem. 65: 1323-1328 (1993).

Oligonucleotide tags of the invention may range in length from 12 to 60 nucleotides or basepairs. Preferably, oligonucleotide tags range in length from 18 to 40 nucleotides or basepairs. More preferably, oligonucleotide tags range in length from 30 to 40 nucleotides or basepairs.

20

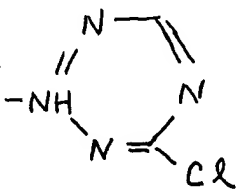
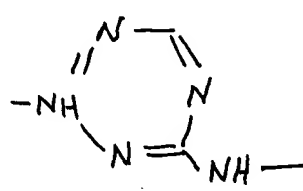
#### Attaching Tags to Molecules

Oligonucleotide tags may be attached to many different classes of molecules by a variety of reactive functionalities well known in the art, e.g. Haugland, Handbook of Fluorescent Probes and Research Chemicals (Molecular  
25 Probes, Inc., Eugene, 1992); Khanna et al, U.S. patent 4,318,846; or the like. Table III provides exemplary functionalities and counterpart reactive groups that may reside on oligonucleotide tags or the molecules of interest. When the functionalities and counterpart reactants are reacted together, after activation in some cases, a linking group is formed.

30

35

Table III  
Reactive Functionalities and Their Counterpart Reactants  
and Resulting Linking Groups

5	Reactive Functionality	Counterpart Functionality	Linking Group
	-NH <sub>2</sub>	-COOH	-CO-NH-
10	-NH <sub>2</sub>	-NCO	-NHCONH-
	-NH <sub>2</sub>	-NCS	-NHCSNH-
15			
20	-NH <sub>2</sub>		
25	-SH	-C=C-CO-	-S-C-C-CO-
	-NH <sub>2</sub>	-CHO	-CH <sub>2</sub> NH-
	-NH <sub>2</sub>	-SO <sub>2</sub> Cl	-SO <sub>2</sub> NH-
30	-OH	-OP(NCH(CH <sub>3</sub> ) <sub>2</sub> ) <sub>2</sub>	-OP(=O)(O)O-
	-OP(=O)(O)S	-NHC(=O)CH <sub>2</sub> Br	-NHC(=O)CH <sub>2</sub> SP(=O)(O)O-

35

A class of molecules particularly convenient for use with the invention includes linear polymeric molecules of the form:



40

wherein L is a linker moiety and M is a monomer that may selected from a wide range of chemical structures to provide a range of functions from serving as an

inert non-sterically hindering spacer moiety to providing a reactive functionality which can serve as a branching point to attach other components, a site for attaching labels; a site for attaching oligonucleotides or other binding polymers for hybridizing or binding to a therapeutic target; or as a site for attaching other  
5 groups for affecting solubility, promotion of duplex and/or triplex formation, such as intercalators, alkylating agents, and the like. The sequence, and therefore composition, of such linear polymeric molecules may be encoded within a polynucleotide attached to the tag, as taught by Brenner and Lerner (cited above). However, after a selection event, instead of amplifying then sequencing the tag of  
10 the selected molecule, the tag can be sequenced directly--using a so-called "single base" approach described below--after releasing the molecule of interest, e.g. by restriction digestion of a site engineered into the tag.

The following references disclose several phosphoramidite and/or hydrogen phosphonate monomers suitable for use in the present invention and provide  
15 guidance for their synthesis and inclusion into oligonucleotides: Newton et al, *Nucleic Acids Research*, 21: 1155-1162 (1993); Griffin et al, *J. Am. Chem. Soc.*, 114: 7976-7982 (1992); Jaschke et al, *Tetrahedron Letters*, 34: 301-304 (1992); Ma et al, International application PCT/CA92/00423; Zon et al, International application PCT/US90/06630; Durand et al, *Nucleic Acids Research*, 18: 6353-  
20 6359 (1990); Salunkhe et al, *J. Am. Chem. Soc.*, 114: 8768-8772 (1992); Urdea et al, U.S. patent 5,093,232; Ruth, U.S. patent 4,948,882; Cruickshank, U.S. patent 5,091,519; Haralambidis et al, *Nucleic Acids Research*, 15: 4857-4876 (1987); and the like. More particularly, M is a straight chain, cyclic, or branched organic molecular structure containing from 1 to 20 carbon atoms and from 0 to 10  
25 heteroatoms selected from the group consisting of oxygen, nitrogen, and sulfur. Preferably, M is alkyl, alkoxy, alkenyl, or aryl containing from 1 to 16 carbon atoms; heterocyclic having from 3 to 8 carbon atoms and from 1 to 3 heteroatoms selected from the group consisting of oxygen, nitrogen, and sulfur; glycosyl; or nucleosidyl. More preferably, M is alkyl, alkoxy, alkenyl, or aryl containing from 1  
30 to 8 carbon atoms; glycosyl; or nucleosidyl.

Preferably, L is a phosphorus(V) linking group which may be phosphodiester, phosphotriester, methyl or ethyl phosphonate, phosphorothioate, phosphorodithioate, phosphoramidate, or the like. Generally, linkages derived from phosphoramidite or hydrogen phosphonate precursors are preferred so that  
35 the linear polymeric units of the invention can be conveniently synthesized with

commercial automated DNA synthesizers, e.g. Applied Biosystems, Inc. (Foster City, CA) model 394, or the like.

5 n may vary significantly depending on the nature of M and L. Usually, n varies from about 3 to about 100. When M is a nucleoside or analog thereof or a nucleoside-sized monomer and L is a phosphorus(V) linkage, then n varies from about 12 to about 100. Preferably, when M is a nucleoside or analog thereof or a nucleoside-sized monomer and L is a phosphorus(V) linkage, then n varies from about 12 to about 40.

10 Peptides are another preferred class of molecules to which tags of the invention are attached. Synthesis of peptide-oligonucleotide conjugates for use in the invention is taught in Haralambidis et al (cited above) and International patent application PCT/AU88/004417; Truffert et al, Tetrahedron Letters, 35: 2353-2356 (1994); de la Torre et al, Tetrahedron Letters, 35: 2733-2736 (1994); and like references.

15

#### Solid Phase Supports

A wide variety of solid phase supports may be employed for use with the invention, including microparticles, beads, and membranes of various materials, glass slides, silicon chips, polystyrene beads,  
20 alkanethiolate-derivatized gold, and the like. A population of discrete particles may be employed such that each has a uniform coating, or population, of the complementary sequence of the same tag, or a single or a few supports may be employed with spacially discrete regions each containing a uniform coating, or population, of complementary sequences  
25 to the same tag.

Tag complements may be used with the solid phase support that they are synthesized on, or they may be separately synthesized and attached to a solid phase support for use, e.g. as disclosed by Lund et al, Nucleic Acids Research, 16: 10861-10880 (1988); Albretsen et al, Anal. Biochem., 189:  
30 40-50 (1990); Wolf et al, Nucleic Acids Research, 15: 2911-2926 (1987); or Ghosh et al, Nucleic Acids Research, 15: 5353-5372 (1987). Preferably, tag complements are synthesized on and used with the same solid phase support, which may comprise a variety of forms and include a variety of linking moieties. Such supports may comprise microparticles or arrays, or  
35 matrices, of regions where uniform populations of tag complements are synthesized. Microparticle supports include commercially available

nucleoside-derivatized CPG and polystyrene (Applied Biosystems, Foster City, CA); derivatized magnetic beads; polystyrene grafted with polyethylene glycol (e.g., TentaGel™, Rapp Polymere, Tübingen Germany); and the like. Selection of the support characteristics, such as material, porosity, size, shape, and the like, and the type of linking moiety employed depends on the conditions under which the tags are used. For example, in applications involving successive processing with enzymes, supports and linkers that minimize steric hinderance of the enzymes and that facilitate access to substrate are preferred. Exemplary linking moieties are disclosed in Pon et al, *Biotechniques*, 6:768-775 (1988); Webb, U.S. patent 4,659,774; Barany et al, International patent application PCT/US91/06103; Brown et al, *J. Chem. Soc. Commun.*, 1989: 891-893; Damha et al, *Nucleic Acids Research*, 18: 3813-3821 (1990); Beattie et al, *Clinical Chemistry*, 39: 719-722 (1993); Maskos and Southern, *Nucleic Acids Research*, 20: 1679-1684 (1992); and the like.

As mentioned above, tag complements may also be synthesized on a single (or a few) solid phase support to form an array of regions uniformly coated with tag complements. That is, within each region in such an array the same tag complement is synthesized. Techniques for synthesizing such arrays are disclosed in McGall et al, International application PCT/US93/03767; Pease et al, *Proc. Natl. Acad. Sci.*, 91: 5022-5026 (1994); Southern and Maskos, International application PCT/GB89/01114; Maskos and Southern (cited above); Southern et al, *Genomics*, 13: 1008-1017 (1992); and Maskos and Southern, *Nucleic Acids Research*, 21: 4663-4669 (1993).

Preferably, the invention is implemented with microparticles or beads uniformly coated with complements of the same tag sequence. Microparticle supports and methods of covalently or noncovalently linking oligonucleotides to their surfaces are well known, as exemplified by the following references: Beaucage and Iyer (cited above); Gait, editor, *Oligonucleotide Synthesis: A Practical Approach* (IRL Press, Oxford, 1984); and the references cited above. Generally, the size and shape of a microparticle is not critical; however, microparticles in the size range of a few, e.g. 1-2, to several hundred, e.g. 200-1000  $\mu\text{m}$  diameter are preferable, as they facilitate the construction and manipulation of large repertoires of oligonucleotide tags with minimal reagent and sample usage.



Preferably, commercially available controlled-pore glass (CPG) or polystyrene supports are employed as solid phase supports in the invention. Such supports come available with base-labile linkers and initial nucleosides attached, e.g. Applied Biosystems (Foster City, CA).

5 Preferably, microparticles having pore size between 500 and 1000  
angstroms are employed.

## Attaching Target Polynucleotides to Microparticles

An important aspect of the invention is the sorting of populations of identical polynucleotides, e.g. from a cDNA library, and their attachment to microparticles or separate regions of a solid phase support such that each microparticle or region has only a single kind of polynucleotide. This latter condition can be essentially met by ligating a repertoire of tags to a population of polynucleotides followed by cloning and sampling of the ligated sequences. A repertoire of oligonucleotide tags can be ligated to a population of polynucleotides in a number of ways, such as through direct enzymatic ligation, amplification using primers containing the tag sequences, and the like. The initial ligating step produces a varied population of tag-polynucleotide conjugates such that a single tag could be attached to many different polynucleotides. However, by taking a sufficiently small sample of the conjugates, the probability of obtaining "doubles," i.e. the same tag on two different polynucleotides, can be made negligible. The probability of obtaining a sample free of doubles can be estimated by a Poisson distribution (with a density  $\lambda$  approximately equal to  $np$ , where  $n$  is number of conjugates in the sample and  $p$  is the probability of successfully selecting a non-double), provided that the "sufficiently small" sample is on the order of several hundred to thousands of conjugates and that only a small fraction of the total tag repertoire is represented in the sample.

30 Preferably, when the population of polynucleotides is messenger RNA (mRNA), oligonucleotides tags are attached by reverse transcribing the mRNA with a set of primers containing complements of tag sequences. An exemplary set of such primers could have the following sequence:

35            5'-mRNA- [A]<sub>n</sub> -3'  
                               [T]<sub>19</sub>GG[W,W,W,C]<sub>9</sub>ACCAGCTGATC-5'-biotin

where "[W,W,W,C]<sub>9</sub>" represents the sequence of an oligonucleotide tag of nine subunits of four nucleotides each and "[W,W,W,C]" represents the subunit sequences listed above, i.e. "W" represents T or A. The underlined sequences identify an optional restriction endonuclease site that can be used to release the polynucleotide from attachment to a solid phase support via the biotin, if one is employed. For the above primer, the complement attached to a microparticle could have the form:

5'-[G,W,W,W]<sub>9</sub>TGG-linker-microparticle

After reverse transcription, the mRNA is removed, e.g. by RNase H digestion, and the second strand of the cDNA is synthesized using, for example, a primer of the following form:

5'-NRRGATCYNNN-3'

where N is any one of A, T, G, or C; R is a purine-containing nucleotide, and Y is a pyrimidine-containing nucleotide. This particular primer creates a Bst Y1 restriction site in the resulting double stranded DNA which, together with the Sal I site, facilitates cloning into a vector with, for example, Bam HI and Xho I sites. After Bst Y1 and Sal I digestion, the exemplary conjugate would have the form:

5'-RCGACCA[C,W,W,W]<sub>9</sub>GG[T]<sub>19</sub>- cDNA -NNNR  
GGT[G,W,W,W]<sub>9</sub>CC[A]<sub>19</sub>- rDNA -NNNYCTAG-5'

Preferably, when the ligase-based method of sequencing is employed, the Bst Y1 and Sal I digested fragments are cloned into a Bam HI-/Xho I-digested vector having the following single-copy restriction sites:

5'-GAGGATGCCTTTATGGATCCACTCGAGATCCCAATCCA-3'  
FokI BamHI XhoI

This adds the Fok I site which will allow initiation of the sequencing process discussed more fully below.

5 A general method for exposing the single stranded tag after amplification involves digesting a target polynucleotide-containing conjugate with the 5'->3' exonuclease activity of T4 DNA polymerase, or a like enzyme. When used in the presence of a single nucleoside triphosphate, such a polymerase will cleave nucleotides from 3' recessed ends present on the non-template strand of a double stranded fragment until a complement of the single nucleoside triphosphate is reached on the  
10 template strand. When such a nucleotide is reached the 5'->3' digestion effectively ceases, as the polymerase's extension activity adds nucleotides at a higher rate than the excision activity removes nucleotides. Consequently, tags constructed with three nucleotides are readily prepared for loading onto solid phase supports.

15 The technique may also be used to preferentially methylate interior Fok I sites of a target polynucleotide while leaving a single Fok I site at the terminus of the polynucleotide unmethylated. First, the terminal Fok I site is rendered single stranded using a polymerase with deoxycytidine triphosphate. The double stranded portion of the fragment is then  
20 methylated, after which the single stranded terminus is filled in with a DNA polymerase in the presence of all four nucleoside triphosphates, thereby regenerating the Fok I site.

After the oligonucleotide tags are prepared for specific hybridization, e.g. by rendering them single stranded as described above, the  
25 polynucleotides are mixed with microparticles containing the complementary sequences of the tags under conditions that favor the formation of perfectly matched duplexes between the tags and their complements. There is extensive guidance in the literature for creating these conditions. Exemplary references providing such guidance include  
30 Wetmur, Critical Reviews in Biochemistry and Molecular Biology, 26: 227-259 (1991); Sambrook et al, Molecular Cloning: A Laboratory Manual, 2nd Edition (Cold Spring Harbor Laboratory, New York, 1989); and the like. Preferably, the hybridization conditions are sufficiently stringent so that only perfectly matched sequences form stable duplexes. Under such  
35 conditions the polynucleotides specifically hybridized through their tags are

ligated to the complementary sequences attached to the microparticles. Finally, the microparticles are washed to remove unligated polynucleotides.

When CPG microparticles conventionally employed as synthesis supports are used, the density of tag complements on the microparticle surface is typically greater than that necessary for some sequencing operations. That is, in sequencing approaches that require successive treatment of the attached polynucleotides with a variety of enzymes, densely spaced polynucleotides may tend to inhibit access of the relatively bulky enzymes to the polynucleotides. In such cases, the polynucleotides are preferably mixed with the microparticles so that tag complements are present in significant excess, e.g. from 10:1 to 100:1, or greater, over the polynucleotides. This ensures that the density of polynucleotides on the microparticle surface will not be so high as to inhibit enzyme access. Preferably, the average inter-polynucleotide spacing on the microparticle surface is on the order of 30-100 nm. Guidance in selecting ratios for standard CPG supports and Ballotini beads (a type of solid glass support) is found in Maskos and Southern, Nucleic Acids Research, 20: 1679-1684 (1992). Preferably, for sequencing applications, standard CPG beads of diameter in the range of 20-50  $\mu\text{m}$  are loaded with about  $10^5$  polynucleotides.

#### Single Base DNA Sequencing

The present invention can be employed with conventional methods of DNA sequencing, e.g. as disclosed by Hultman et al, Nucleic Acids Research, 17: 4937-4946 (1989). However, for parallel, or simultaneous, sequencing of multiple polynucleotides, a DNA sequencing methodology is preferred that does not require electrophoretic separation of closely sized DNA fragments. Several such so-called "single base" methods are available, which can be adapted to parallel operation when combined with the present invention. Single base approaches are disclosed in the following references: U.S. patent 5,302,509; and U.S. patent 4,971,903.

A "single base" method of DNA sequencing which is suitable for use with the present invention and which requires no electrophoretic separation of DNA fragments is described in co-pending U.S. patent application 08/280,441 filed 25 July 1994, which application is incorporated by reference. The method comprises the following steps: (a) ligating a probe

to an end of the polynucleotide having a protruding strand to form a ligated complex, the probe having a complementary protruding strand to that of the polynucleotide and the probe having a nuclease recognition site; (b) removing unligated probe from the ligated complex; (c) identifying one or more nucleotides in the protruding strand of the polynucleotide by the identity of the ligated probe; (d) cleaving the ligated complex with a nuclease; and (e) repeating steps (a) through (d) until the nucleotide sequence of the polynucleotide is determined. As is described more fully in the above cited patent application, identifying the one or more nucleotides can be carried out either before or after cleavage of the ligated complex from the target polynucleotide. Preferably, whenever natural protein endonucleases are employed, the method further includes a step of methylating the target polynucleotide at the start of a sequencing operation.

15                   Apparatus for Observing Enzymatic Processes and/or  
                          Binding Events at Microparticle Surfaces

An objective of the invention is to sort identical molecules, particularly polynucleotides, onto the surfaces of microparticles by the specific hybridization of tags and their complements. Once such sorting has taken place, the presence of the molecules or operations performed on them can be detected in a number of ways depending on the nature of the tagged molecule, whether microparticles are detected separately or in "batches," whether repeated measurements are desired, and the like. Typically, the sorted molecules are exposed to ligands for binding, e.g. in drug development, or are subjected chemical or enzymatic processes, e.g. in polynucleotide sequencing. In both of these uses it is often desirable to simultaneously observe signals corresponding to such events or processes on large numbers of microparticles. Microparticles carrying sorted molecules (referred to herein as "loaded" microparticles) lend themselves to such large scale parallel operations, e.g. as demonstrated by Lam et al (cited above).

Preferably, whenever light-generating signals, e.g. chemiluminescent, fluorescent, or the like, are employed to detect events or processes, loaded microparticles are spread on a planar substrate, e.g. a glass slide, for examination with a scanning system, such as described in International patent application PCT/US91/09217. The scanning system

should be able to reproducibly scan the substrate and to define the positions of each microparticle in a predetermined region by way of a coordinate system. In polynucleotide sequencing applications, it is important that the positional identification of microparticles be repeatable in successive scan steps.

Such scanning systems may be constructed from commercially available components, e.g. x-y translation table controlled by a digital computer used with a detection system consisting of a photomultiplier tube and appropriate optics, e.g. for exciting and collecting fluorescent signals.

The stability and reproducibility of the positional localization in scanning will determine, to a large extent, the resolution for separating closely spaced microparticles. Preferably, the scanning systems should be capable of resolving closely spaced microparticles, e.g. separated by a particle diameter. Thus, for most applications, e.g. using CPG microparticles, the scanning system should at least have the capability of resolving objects on the order of 10-100  $\mu\text{m}$ . Even higher resolution may be desirable in some embodiments, but with increase resolution, the time required to fully scan a substrate will increase; thus, in some embodiments a compromise may have to be made between speed and resolution. Increases in scanning time can be achieved by a system which only scans positions where microparticles are known to be located, e.g. from an initial full scan.

In sequencing applications, loaded microparticles can be fixed to the surface of a substrate in variety of ways. Preferably, when the substrate is glass, its surface is derivatized with an alkylamino linker using commercially available reagents, e.g. Pierce Chemical, which in turn is cross-linked to avidin, again using conventional chemistries, to form an avidinated surface. Biotin moieties can be introduced to the loaded microparticles in a number of ways. For example, a fraction, e.g. 10-15 percent, of the cloning vectors used to attach tags to polynucleotides are engineered to contain a unique restriction site (providing sticky ends on digestion) immediately adjacent to the polynucleotide insert at an end of the polynucleotide opposite of the tag. The site is excised with the polynucleotide and tag for loading onto microparticles. After loading, about 10-15 percent of the loaded polynucleotides will possess the unique restriction site distal from the microparticle surface. After digestion with the associated restriction endonuclease, an appropriate double stranded adaptor containing a biotin

moiety is ligated to the sticky end. The resulting microparticles are then spread on the avidinated glass surface where they become fixed via the biotin-avidin linkages.

Alternatively and preferably when sequencing by ligation is employed, in the initial ligation step a mixture of probes is applied to the loaded microparticle: a fraction of the probes contain a type II's restriction recognition site, as required by the sequencing method, and a fraction of the probes have no such recognition site, but instead contain a biotin moiety at its non-ligating end. Preferably, the mixture comprises about 10-15 percent of the biotinylated probe.

#### Parallel Sequencing

The tagging system of the invention can be used with the preferred single base sequencing method to sequence polynucleotides up to several kilobases in length. The tagging system permits many thousands of randomly overlapping fragments of a target polynucleotide to be sequenced simultaneously. Sequencing proceeds in a stepwise fashion at each of the many thousands of microparticles loaded with distinct fragments and fixed to a common substrate associated with a scanning system, such as that described above. Preferably, from 20-50 bases are identified at each microparticle. With this information, the sequence of the target polynucleotide is determined by collating the 20-50 base fragments via their overlapping regions, e.g. as described in U.S. patent 5,002,867.

Fragments are generated from a target polynucleotide following protocols employed in conventional "shotgun" sequencing, e.g. as disclosed in Sambrook et al (cited above). Briefly, starting with a target polynucleotide as an insert in an appropriate cloning vector, e.g.  $\lambda$  phage, the vector is expanded, purified and digested with the appropriate restriction enzymes to yield about 10-15  $\mu$ g of purified insert. Typically, the protocol results in about 500-1000 subclones per microgram of starting DNA. The insert is separated from the vector fragments by preparative gel electrophoresis, removed from the gel by conventional methods, and resuspended in a standard buffer, such as TE (Tris-EDTA). The restriction enzymes selected to excise the insert from the vector preferably leave compatible sticky ends on the insert, so that the insert can be self-ligated in preparation for generating randomly overlapping fragments. As explained

in Sambrook et al (cited above), the circularized DNA yields a better random distribution of fragments than linear DNA in the fragmentation methods employed below. After self-ligating the insert, e.g. with T4 ligase using conventional protocols, the purified ligated insert is fragmented by a standard protocol, e.g. sonication or DNase I digestion in the presence of  $Mn^{++}$ . After fragmentation the ends of the fragments are repaired, e.g. as described in Sambrook et al (cited above), and the repaired fragments are separated by size using gel electrophoresis. Fragments in the 300-500 basepair range are selected and eluted from the gel by conventional means, and ligated into a tag-carrying vector as described above to form a library of tag-fragment conjugates.

As described above, a sample containing several thousand tag-fragment conjugates are taken from the library and expanded, after which the tag-fragment inserts are excised from the vector and prepared for specific hybridization to the tag complements as described above.

### Example I

#### Sorting Multiple Target Polynucleotides Derived from pUC19

A mixture of three target polynucleotide-tag conjugates are obtained as follows: First, the following six oligonucleotides are synthesized and combined pairwise to form tag 1, tag 2, and tag 3:

5'-pTCGACC(w<sub>1</sub>)(w<sub>2</sub>)(w<sub>3</sub>)(w<sub>4</sub>)(w<sub>5</sub>)(w<sub>6</sub>)(w<sub>7</sub>)(w<sub>8</sub>)(w<sub>1</sub>)A  
GG(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)TTCGAp-5'

#### Tag 1

5'-pTCGACC(w<sub>6</sub>)(w<sub>7</sub>)(w<sub>8</sub>)(w<sub>1</sub>)(w<sub>2</sub>)(w<sub>6</sub>)(w<sub>4</sub>)(w<sub>2</sub>)(w<sub>1</sub>)A  
GG(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)TTCGAp-5'

#### Tag 2

5'-pTCGACC(w<sub>3</sub>)(w<sub>2</sub>)(w<sub>1</sub>)(w<sub>1</sub>)(w<sub>5</sub>)(w<sub>8</sub>)(w<sub>8</sub>)(w<sub>4</sub>)(w<sub>4</sub>)A  
GG(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)TTCGAp-5'



Tag 3

where "p" indicates a monophosphate, the  $w_i$ 's represent the subunits define in Table I, and the terms "(\*\*)" represent their respective complements. A  
5 pUC19 is digested with Sal I and Hind III, the large fragment is purified, and separately ligated with tags 1, 2, and 3, to form pUC19-1, pUC19-2, and pUC19-3, respectively. The three recombinants are separately amplified and isolated, after which pUC19-1 is digested with Hind III and Aat I, pUC19-2 is digested with Hind III and Ssp I, and pUC19-3 is  
10 digested with Hind III and Xmn I. The small fragments are isolated using conventional protocols to give three double stranded fragments about 250, 375, and 575 basepairs in length, respectively, and each having a recessed 3' strand adjacent to the tag and a blunt or 3' protruding strand at the opposite end. Approximately 12 nmoles of each fragment are mixed with 5  
15 units T4 DNA polymerase in the manufacturer's recommended reaction buffer containing 33  $\mu$ M deoxycytosine triphosphate. The reaction mixture is allowed to incubate at 37°C for 30 minutes, after which the reaction is stopped by placing on ice. The fragments are then purified by conventional means.

20 CPG microparticles (37-74  $\mu$ m particle size, 500 angstrom pore size, Pierce Chemical) are derivatized with the linker disclosed by Maskos and Southern, Nucleic Acids Research, 20: 1679-1684 (1992). After separating into three aliquots, the complements of tags 1, 2, and 3 are synthesized on the microparticles using a conventional automated DNA synthesizer, e.g. a  
25 model 392 DNA synthesizer (Applied Biosystems, Foster City, CA). Approximately 1 mg of each of the differently derivatized microparticles are placed in separate vessels.

The T4 DNA polymerase-treated fragments excised from pUC19-1, -  
2, and -3 are resuspended in 50  $\mu$ L of the manufacturer's recommended  
30 buffer for Taq DNA ligase (New England Biolabs). The mixture is then equally divided among the three vessels containing the 1 mg each of derivatized CPG microparticles. 5 units of Taq DNA ligase is added to each vessel, after which they are incubated at 55°C for 15 minutes. The reaction is stopped by placing on ice and the microparticles are washed  
35 several times by repeated centrifugation and resuspension in TE. Finally, the microparticles are resuspended in Nde I reaction buffer (New England

Biolabs) where the attached polynucleotides are digested. After separation from the microparticles the polynucleotide fragments released by Nde I digestion are fluorescently labeled by incubating with Sequenase DNA polymerase and fluorescein labeled thymidine triphosphate (Applied Biosystems, Foster City, CA). The fragments are then separately analyzed on a nondenaturing polyacrylamide gel using an Applied Biosystems model 373 DNA sequencer.

## Example II

### Parallel Sequencing of SV40 Fragments

A repertoire of 36-mer tags consisting of nine 4-nucleotide subunits selected from Table I is prepared by separately synthesizing tags and tag complements by a split and mix approach, as described above. The repertoire is synthesized so as to permit ligation into a Sma I/Hind III digested M13mp19. Thus, as in Example I, one set of oligonucleotides begins with the addition of A followed by nine rounds of split and mix synthesis wherein the oligonucleotide is extended subunit-wise by 3'-phosphoramidite derivatized 4-mers corresponding to the subunits of Table I. The synthesis is then completed with the nucleotide-by-nucleotide addition of one half of the Sma I recognition site (GGG), two C's, and a 5'-monophosphate, e.g. via the Phosphate-ON reagent available from Clontech Laboratories (Palo Alto, CA). The other set of oligonucleotides begins with the addition of three C's (portion of the Sma I recognition site) and two G's, followed by nine rounds of split and mix synthesis wherein the oligonucleotide is extended by 3'-phosphoramidite derivatized 4-mers corresponding to the complements of the subunits of Table I. Synthesis is completed by the nucleotide-by-nucleotide addition of the Hind III recognition site and a 5'-monophosphate. After separation from the synthesis supports the oligonucleotides are mixed under conditions that permit formation of the following duplexes:

5'-pGGGCC(w<sub>1</sub>)(w<sub>1</sub>)(w<sub>1</sub>)(w<sub>1</sub>)(w<sub>1</sub>)(w<sub>1</sub>)(w<sub>1</sub>)(w<sub>1</sub>)A  
 CCCGG(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)(\*\*)TTCGAp-5'

The mixture of duplexes is then ligated into a Sma I/Hind III-digested M13mp19. A repertoire of tag complements are synthesized on CPG microparticles as described above.

Next the following adaptor is prepared which contains a Fok I site and portions of Eco RI and Sma I sites:

5' -pAATTCGGATGATGCATGCATCGACCC  
 GCCTACTACGTACGTAGCTGGGp-5'  
 Eco RI      Fok I                      Sma I

The adaptor is ligated into the Eco RI/Sma I digested M13 described above.

Separately, SV40 DNA is fragmented by sonication following the protocol set forth in Sambrook et al (cited above). The resulting fragments are repaired using standard protocols and separated by size. Fragments in the range of 300-500 basepairs are selected and ligated into the Sma I digested M13 described above to form a library of fragment-tag conjugates, which is then amplified. A sample containing several thousand different fragment-tag conjugates is taken from the library, further amplified, and the fragment-tag inserts are excised by digesting with Eco RI and Hind III. The excised fragment-tag conjugates are treated with T4 DNA polymerase in the presence of deoxycytidine triphosphate, as described in Example I, to expose the oligonucleotide tags for specific hybridization to the CPG microparticles.

After hybridization and ligation, as described in Example I, the loaded microparticles are treated with Fok I to produce a 4-nucleotide protruding strand of a predetermined sequence. A 10:1 mixture (probe 1:probe 2) of the following probes are ligated to the polynucleotides on microparticles.

Probe 1                      FAM- ATCGGATGAC  
    TAGCCTACTGAGCT  
 Probe 2                      biotin- ATCGGATGAC  
    TAGCCTACTGAGCT

FAM represents a fluorescein dye attached to the 5'-hydroxyl of the top strand of Probe 1 through an aminophosphate linker available from Applied Biosystems (Aminolinker). The biotin may also be attached through an Aminolinker moiety and optionally may be further extended via polyethylene oxide linkers, e.g. Jaschke et al (cited above).

The loaded microparticles are then deposited on the surface of an avidinated glass slide to which and from which reagents and wash solutions can be delivered and removed. The avidinated slide with the attached microparticles is examined with a scanning fluorescent microscope (e.g. Zeiss Axioskop equipped with a Newport Model PM500-C motion controller, a Spectra-Physics Model 2020 argon ion laser producing a 488 nm excitation beam, and a 520 nm long-pass emission filter, or like apparatus). The excitation beam and fluorescent emissions are delivered and collected, respectively, through the same objective lens. The excitation beam and collected fluorescence are separated by a dichroic mirror which directs the collected fluorescence to a photon-counting device, e.g. comprising a Hamamatsu model 9403-02 photomultiplier, a Stanford Research Systems model SR445 amplifier and model SR430 multichannel scaler, and digital computer, e.g. a 486-based computer. The computer generates a two dimensional map of the slide which registers the positions of the microparticles.

After cleavage with Fok I to remove the initial probe, the polynucleotides on the attached microparticles undergo 20 cycles of probe ligation, washing, detection, cleavage, and washing, in accordance with the preferred single base sequencing methodology described below. Within each detection step, the scanning system records the fluorescent emission corresponding the base identified at each microparticle. Reactions and washes below are generally carried out with manufacturer's (New England Biolabs') recommended buffers for the enzymes employed, unless otherwise indicated. Standard buffers are also described in Sambrook et al (cited above).

The following four sets of mixed probes are provided for addition to the target polynucleotides:

TAMRA- ATCGGATGACATCAAC  
TAGCCTACTGTAGTTGANNN

FAM- ATCGGATGACATCAAC  
TAGCCTACTGTAGTTGCNNN

ROX- ATCGGATGACATCAAC

**TAGCCTACTGTAGTTGGNNN**

**JOE- ATCGGATGACATCAAC**  
**TAGCCTACTGTAGTTGTNNN**

5

where TAMRA, FAM, ROX, and JOE are spectrally resolvable fluorescent labels attached by way of Aminolinker II (all being available from Applied Biosystems, Inc., Foster City, California); the bold faced nucleotides are the recognition site for Fok I endonuclease, and "N" represents any one of the four nucleotides, A, C, G, T. TAMRA (tetramethylrhodamine), FAM (fluorescein), ROX (rhodamine X), and JOE (2',7'-dimethoxy-4',5'-dichlorofluorescein) and their attachment to oligonucleotides is also described in Fung et al, U.S. patent 4,855,225.

15 The above probes are incubated in approximately 5 molar excess of the target polynucleotide ends as follows: the probes are incubated for 60 minutes at 16°C with 200 units of T4 DNA ligase and the anchored target polynucleotide in T4 DNA ligase buffer; after washing, the target polynucleotide is then incubated with 100 units T4 polynucleotide kinase in the manufacturer's recommended buffer for 30 minutes at 37°C, washed,  
20 and again incubated for 30 minutes at 16°C with 200 units of T4 DNA ligase and the anchored target polynucleotide in T4 DNA ligase buffer. Washing is accomplished by successively flowing volumes of wash buffer over the slide, e.g. TE, disclosed in Sambrook et al (cited above). After the cycle of ligation-phosphorylation-ligation and a final washing, the attached  
25 microparticles are scanned for the presence of fluorescent label, the positions and characteristics of which are recorded by the scanning system. The labeled target polynucleotide, i.e. the ligated complex, is then incubated with 10 units of Fok I in the manufacturer's recommended buffer for 30 minutes at 37°C, followed by washing in TE. As a result the target  
30 polynucleotide is shortened by one nucleotide on each strand and is ready for the next cycle of ligation and cleavage. The process is continued until twenty nucleotides are identified.

**APPENDIX I**  
**Exemplary computer program for generating**  
**minimally cross hybridizing sets**

Program minxh

```
C
C
C
      integer*2 sub1(6),mset1(1000,6),mset2(1000,6)
      dimension nbase(6)
C
C
      write(*,*) 'ENTER SUBUNIT LENGTH'
      read(*,100) nsub
100    format(i1)
      open(1,file='sub4.dat',form='formatted',status='new')
C
C
      nset=0
      do 7000 m1=1,3
        do 7000 m2=1,3
          do 7000 m3=1,3
            do 7000 m4=1,3
              sub1(1)=m1
              sub1(2)=m2
              sub1(3)=m3
              sub1(4)=m4
C
C
      ndiff=3
C
C
C
C
C
C
C
      Generate set of subunits differing from
      sub1 by at least ndiff nucleotides.
      Save in mset1.
C
      jj=1
      do 900 j=1,nsub
900    mset1(1,j)=sub1(j)
C
C
      do 1000 k1=1,3
        do 1000 k2=1,3
          do 1000 k3=1,3
            do 1000 k4=1,3
C
C
              nbase(1)=k1
              nbase(2)=k2
              nbase(3)=k3
```



C	10	C
C	10	C
	15	
C	C	
C	C	
C	C	
C	C	
C	C	
C	C	
C	C	
	20	
		C
		C
	70	
	70	
	70	
	13	C
	70	C